High Speed Wide Area Data Movement: Challenges in the era of 10+ Gbit/s Networks

Raj Kettimuthu

Argonne National Laboratory and

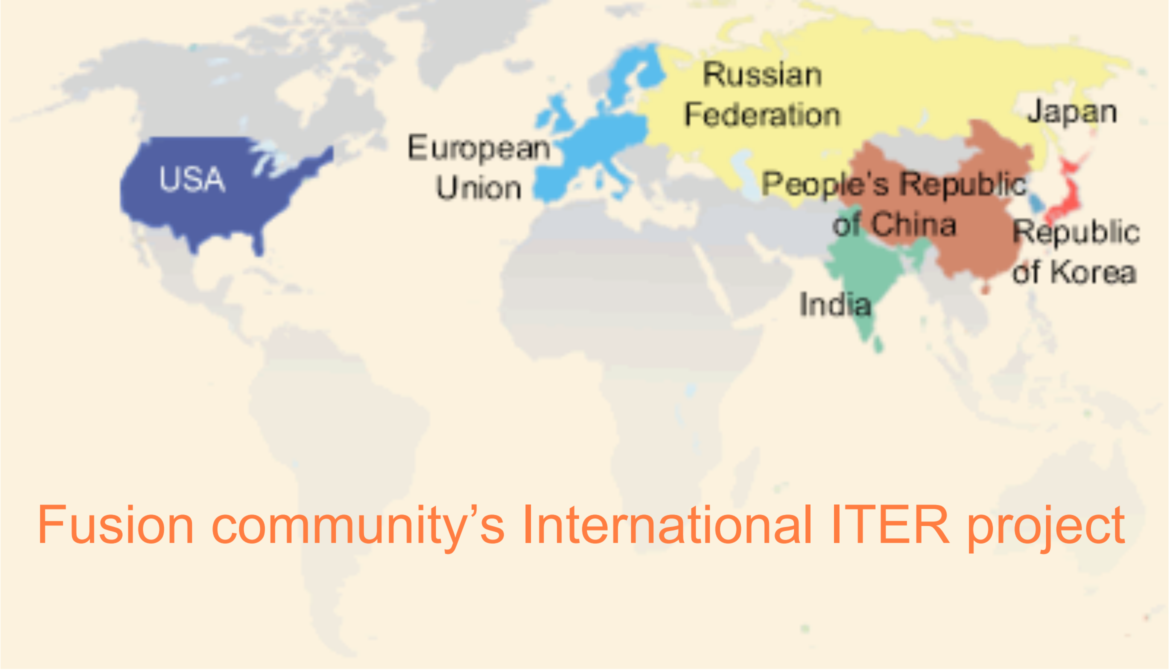
The University of Chicago

Outline

- Introduction
- Network Capabilities
- End-to-End Problem
- GridFTP
- Challenges in 10+ Gbit/s Networks
- Globus.org – Hosted Data Movement Service

Today’s Science Environments

- Large-scale collaborative science is becoming increasingly common



Fusion community’s International ITER project

- Distributed community of users to access and analyze large amounts of data

Simulation Science

- In simulation science, the data sources are supercomputer simulations
  - For eg, climate modeling groups generate large reference simulations at supercomputer centers
  - Many climate scientists need to extract and analyze subsets of this data in various ways
- Combustion, fusion, computational chemistry, and astrophysics communities have similar requirements for remote and distributed data analysis

Experimental Science

- Data sources are facilities such as high energy and nuclear physics experiments and light sources.
  - For eg, LHC at CERN will produce petabytes of raw data per year for 15 years
  - Thousands of physicists worldwide will participate in the analysis
- DOE light sources can also produce large quantities of data that must be distributed, analyzed, and visualized
- The international fusion experiment, ITER

Science Environments

- Raw simulation or observational data is just a starting point for most investigations
- Understanding comes from further analysis, reduction, visualization, and exploration

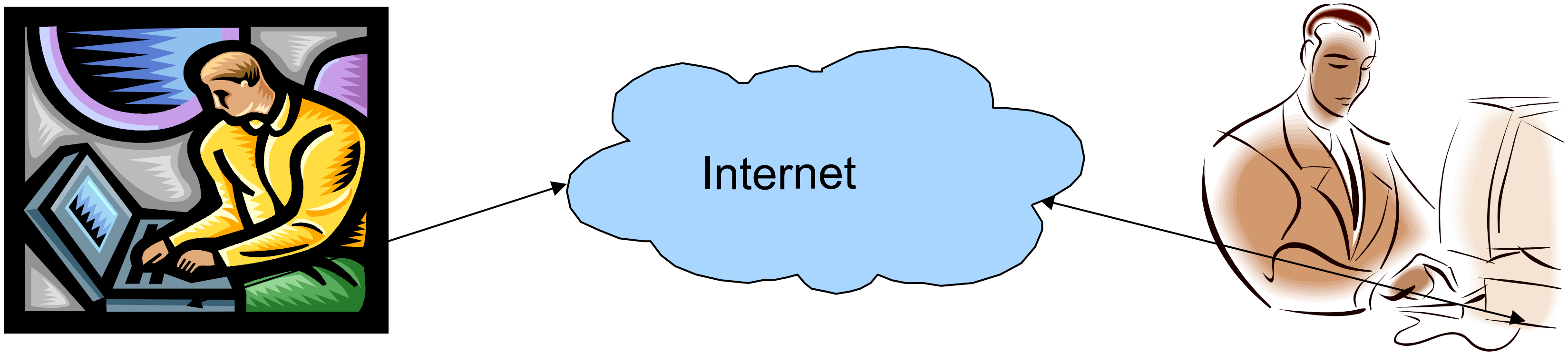
Petascale resource

Compute Cluster

Scientist’s Desktop

- Furthermore the data is a community asset that must be accessible to any member of a distributed collaboration

Network Capabilities



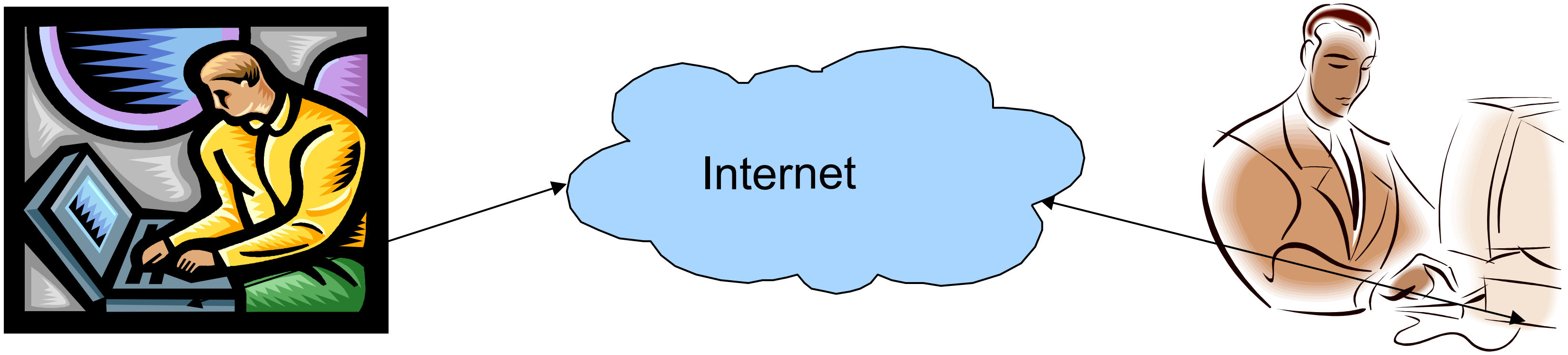
Internet

Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 10 Terabytes of data to Scientist B
- What is the fastest way to transfer the data?

Network Capabilities



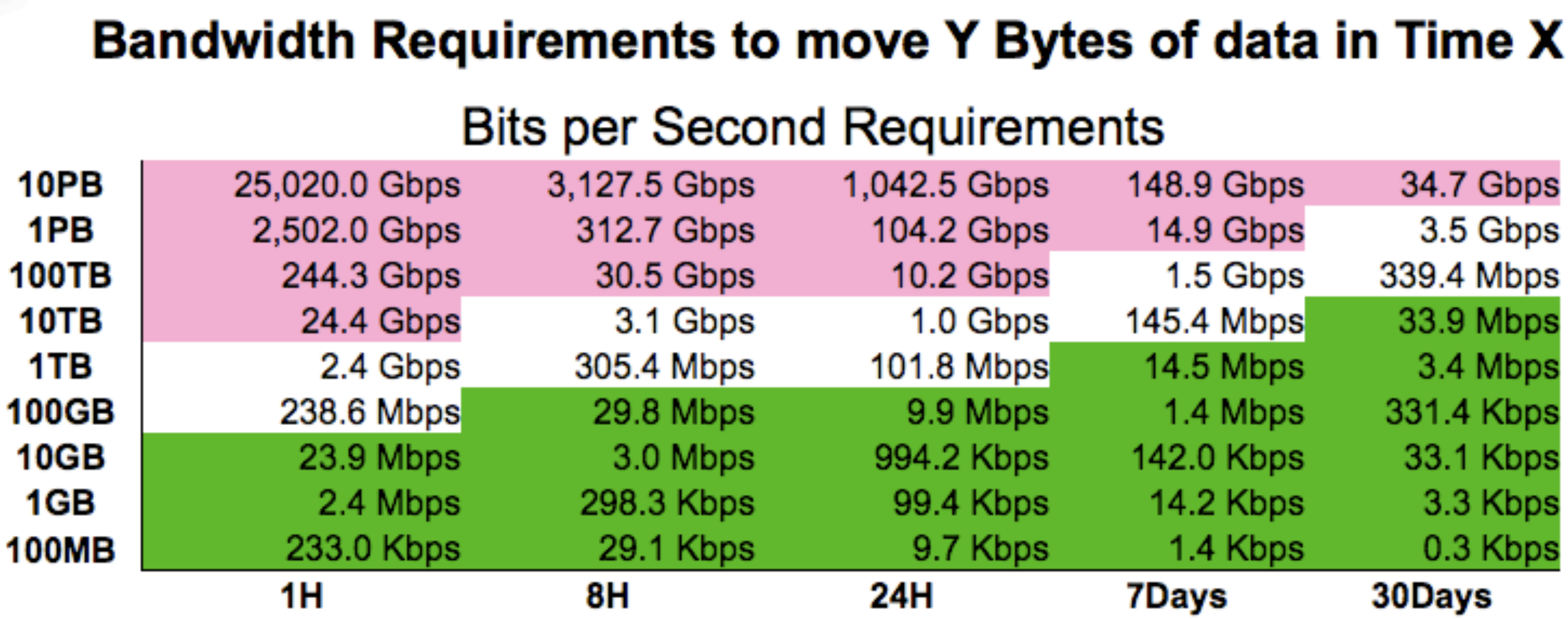
Scientist A in California

Scientist B in New York

- Scientist A wants to transfer 10 Terabytes of data to Scientist B
- What is the fastest way to transfer the data?

FedEx

Bandwidth Requirements

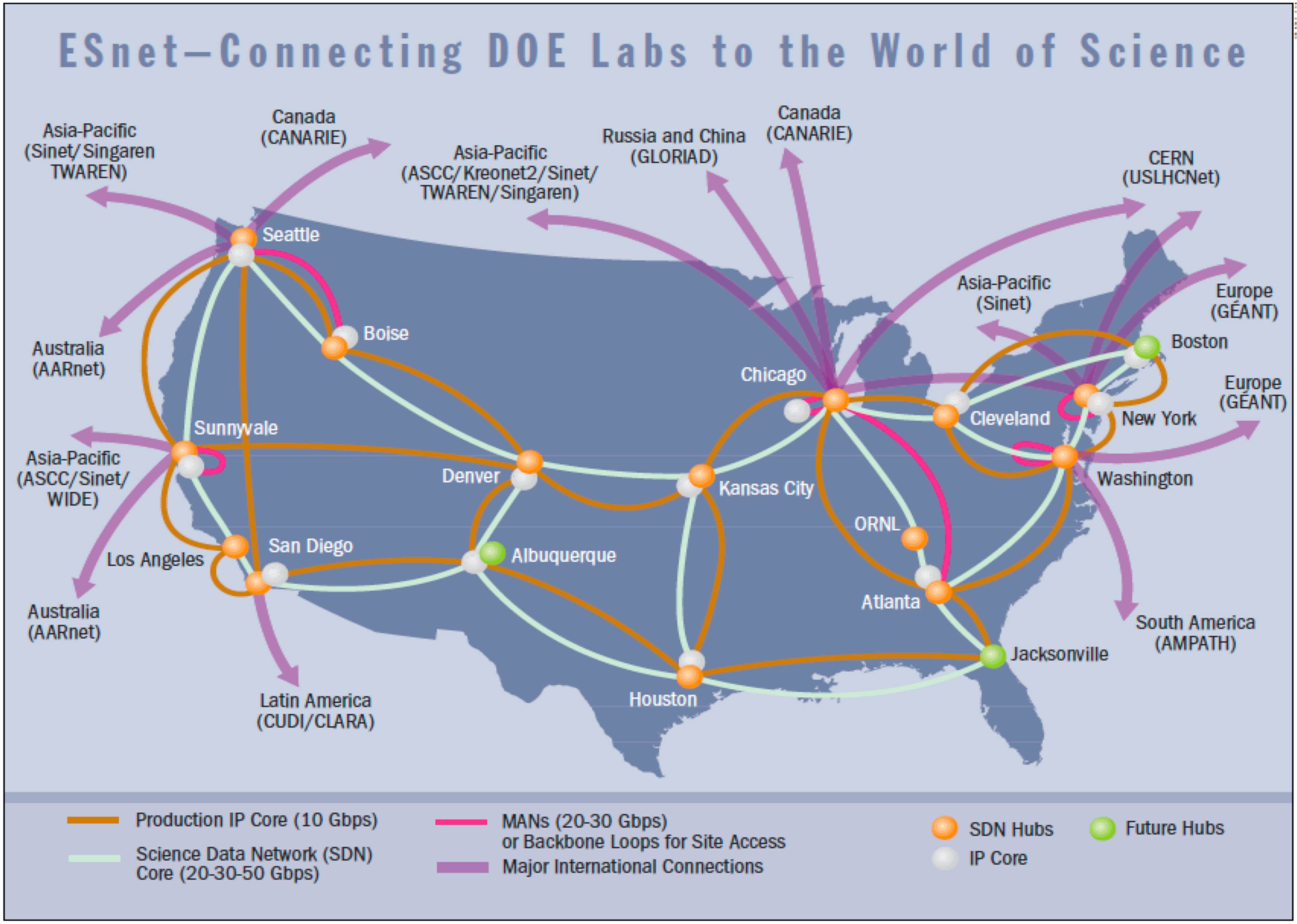


Bandwidth Requirements to move Y Bytes of data in Time X

Bits per Second Requirements

	1H	8H	24H	7Days	30Days
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps

ESNET


ESnet—Connecting DOE Labs to the World of Science
Asia-Pacific (Sinet/ Singaren TWAREN)
Canada (CANARIE)
Asia-Pacific (ASCC/Kreonet2/Sinet/ TWAREN/Singaren)
Russia and China (GLORIAD)
Canada (CANARIE)
CERN (USLHCNet)
Asia-Pacific (Sinet)
Europe (GÉANT)
Boston
Europe (GÉANT)
Seattle
Boise
Australia (AARnet)
Chicago
Cleveland
New York
Sunnyvale
Washington
Asia-Pacific (ASCC/Sinet/ WIDE)
Denver
Kansas City
ORNL
Los Angeles
San Diego
Albuquerque
Atlanta
Australia (AARnet)
South America (AMPATH)
Jacksonville
Latin America (CUDI/CLARA)
Houston
Production IP Core (10 Gbps)
Science Data Network (SDN) Core (20-30-50 Gbps)
MANs (20-30 Gbps) or Backbone Loops for Site Access
Major International Connections
SDN Hubs
IP Core
Future Hubs

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross the globe in less than a second

End-to-end problem

- Now that high-speed networks are available, can we move data at network speeds on the network?
- What if the speed of airplanes had increased by the same factor as computers over the last 50 years, namely five orders of magnitude?

We would be able to cross the globe in less than a second

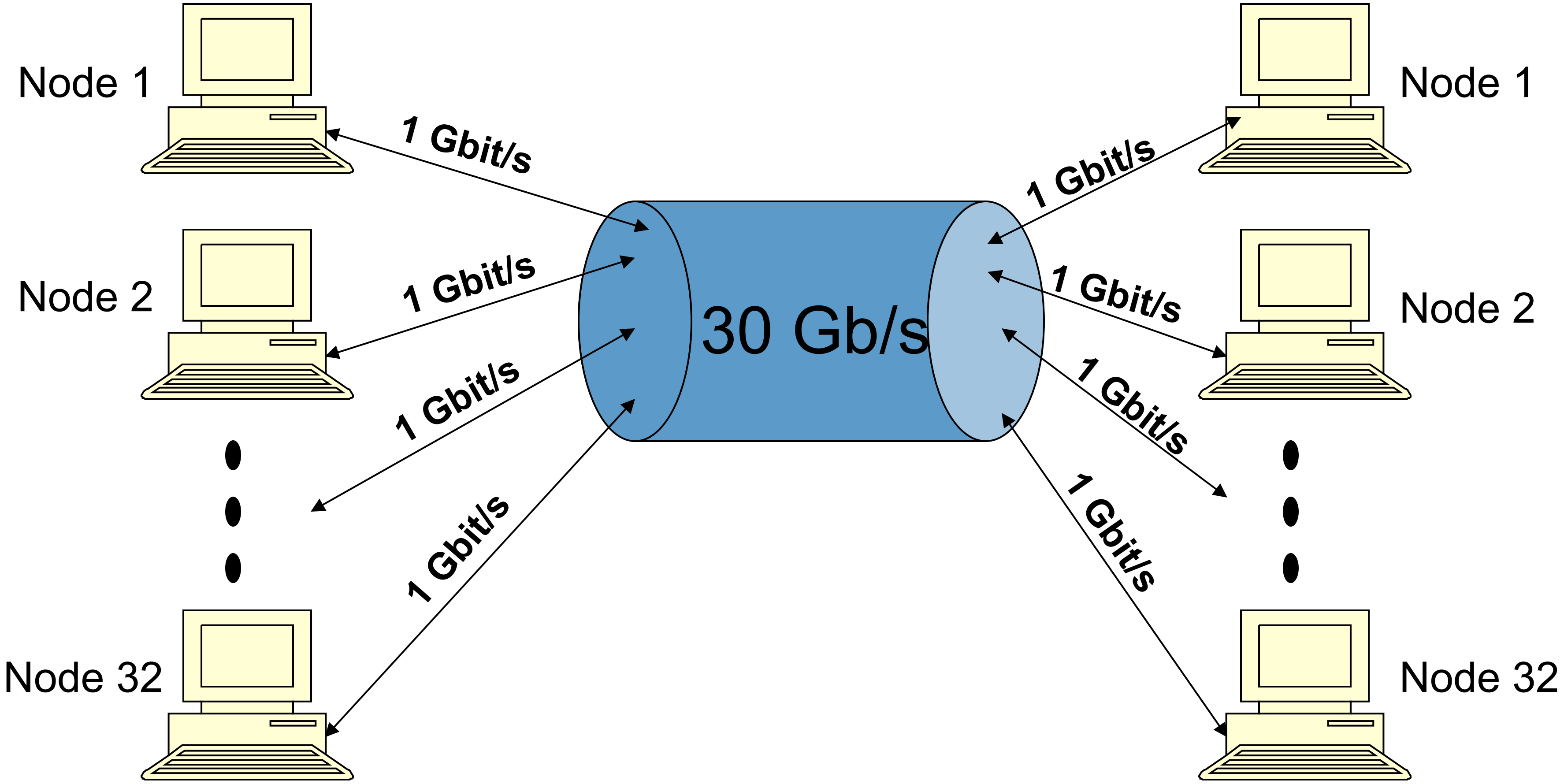
Yes. But it would still take two hours to get to downtown

End-to-end problem

- Data movement in distributed science environments is an end-to-end problem
- A 10 Gbit/s network link between the source and destination does not guarantee an end-to-end data rate of 10 Gbit/s
- Other factors such as storage system, disk, data rate supported by the end node
- Deal with failures of various sorts
  - Firewalls can cause difficulties

End-to-end data transfer

Efficient and robust wide area data transport requires the management of complex systems at multiple levels.



San Diego, CA

Chicago, IL

Challenges

- Standard
- Throughput
- Robustness
- Secure
- Scalable
- Extensible
- Reliable

GridFTP

- High-performance, reliable data transfer protocol optimized for high-bandwidth wide-area networks
- Based on FTP protocol - defines extensions for high-performance operation and security
- Standardized through Open Grid Forum (OGF)
- GridFTP is the OGF recommended data movement protocol

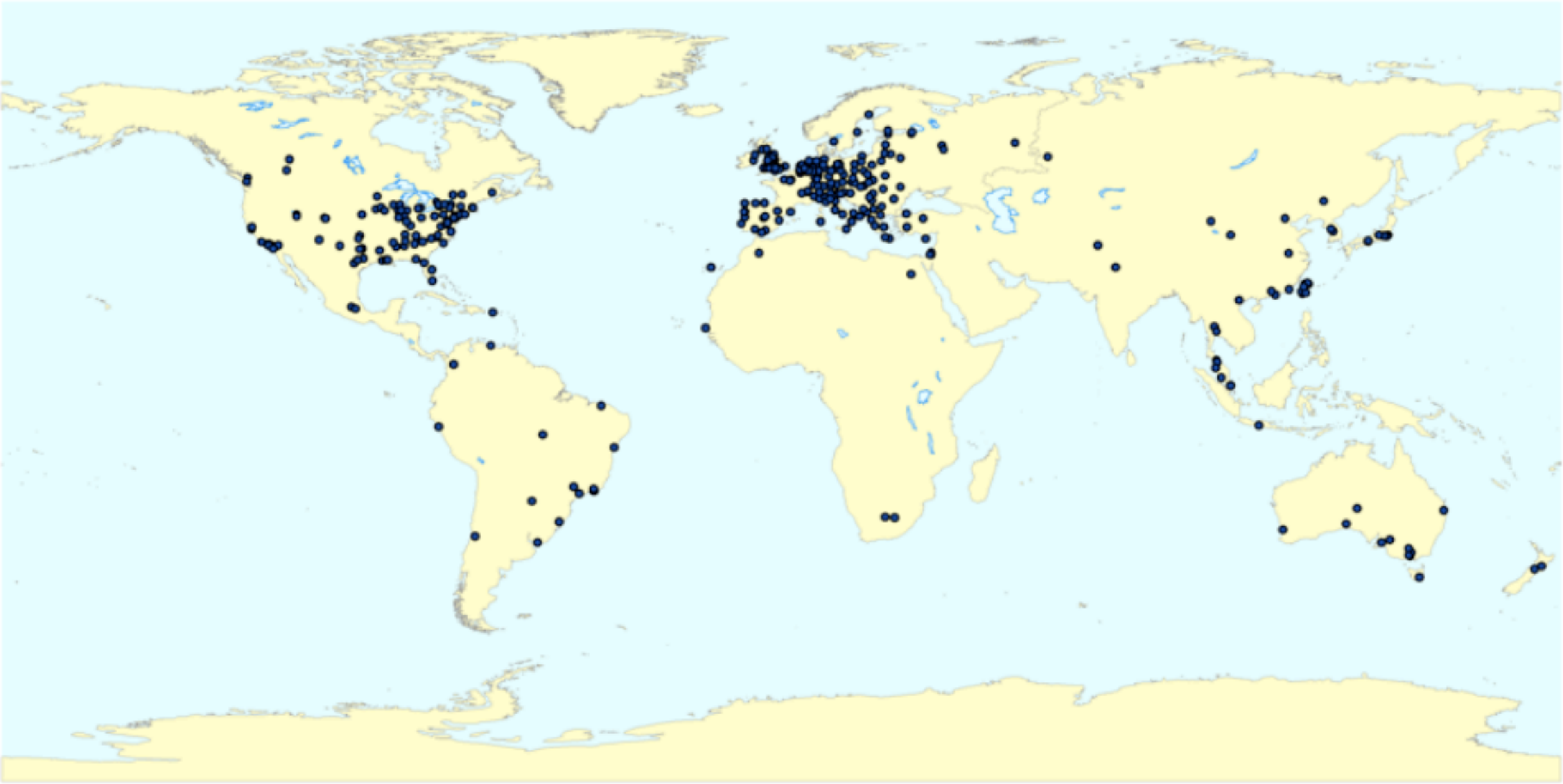
GridFTP

- We (Globus Alliance) supply a reference implementation:
  - Server
  - Client tools
  - Development Libraries
- Multiple independent implementations can interoperate
  - Fermi Lab and U. Virginia have home grown servers that work with ours

Globus GridFTP

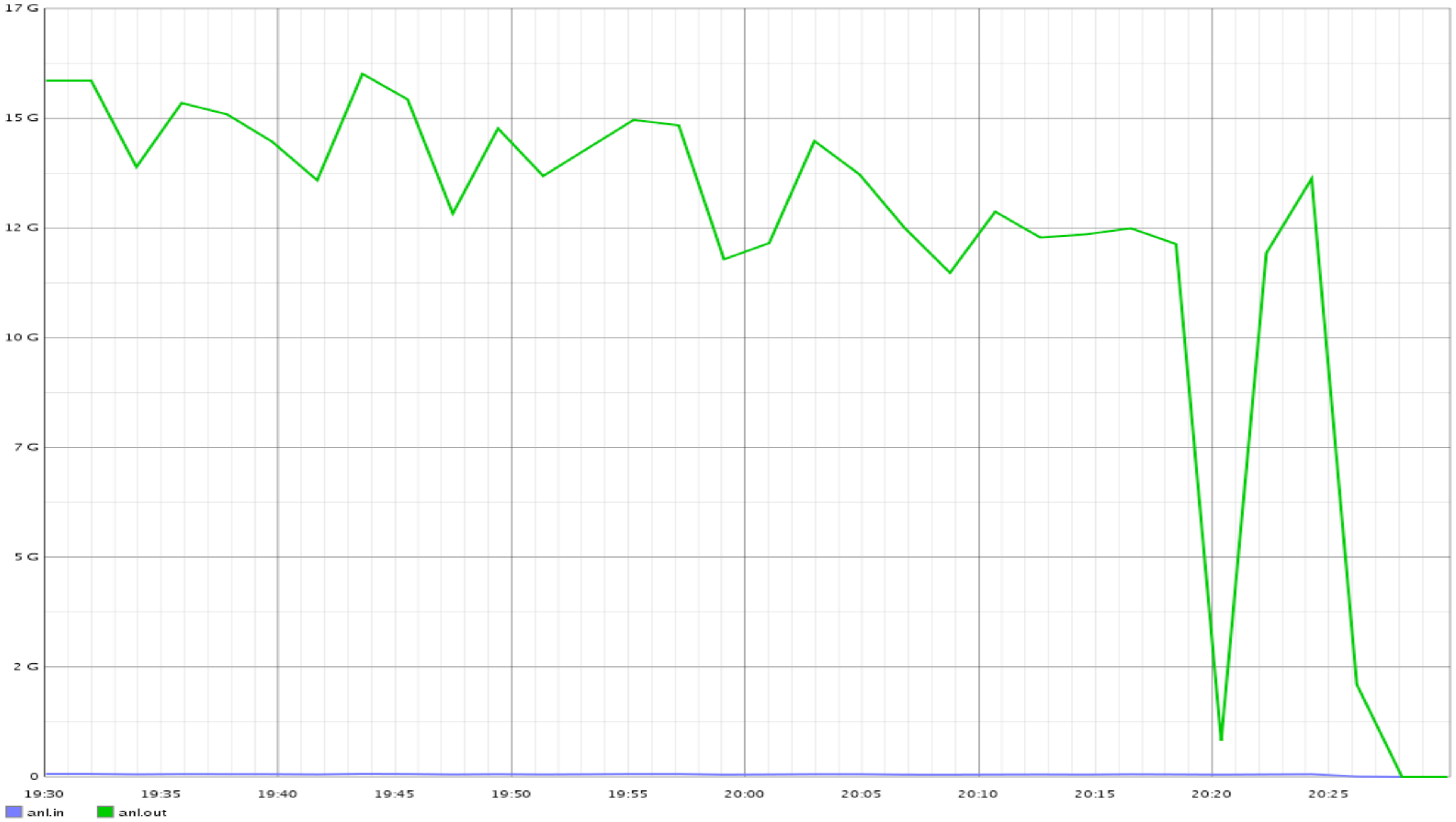
- Performance
  - Parallel TCP streams, optimal TCP buffer
  - Non TCP protocol such as UDT
- Cluster-to-cluster data movement
- Multiple security options
  - Anonymous, password, SSH, GSI
- Support for reliable and restartable transfers

GridFTP Servers Around the World



Created by Tim Pinkawa (Northern Illinois University) using MaxMind's GeoIP technology (http://www.maxmind.com/app/ip-locate).

Performance

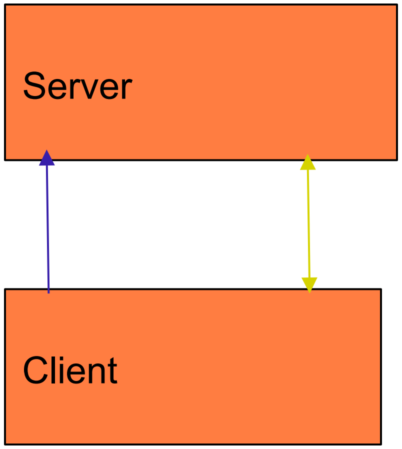


Understanding GridFTP

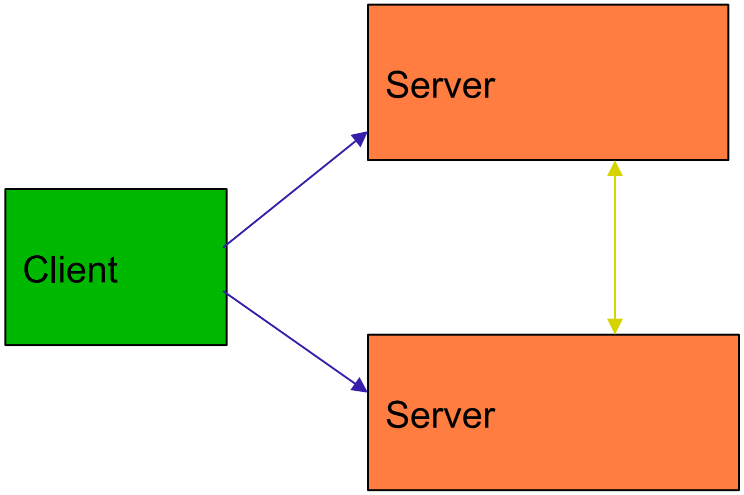
- Two channel protocol like FTP
- Control Channel
  - Command/Response
  - Used to establish data channels
  - Basic file system operations eg. mkdir, delete etc
- Data channel
  - Pathway over which *file* is transferred
  - Many different underlying protocols can be used
    - MODE command determines the protocol

Client/Server and 3rd Party Transfers

- Two party transfer
  - The client connects and forms a CC with the server
  - Information is exchanged to establish the DC
  - A file is transferred over the DC



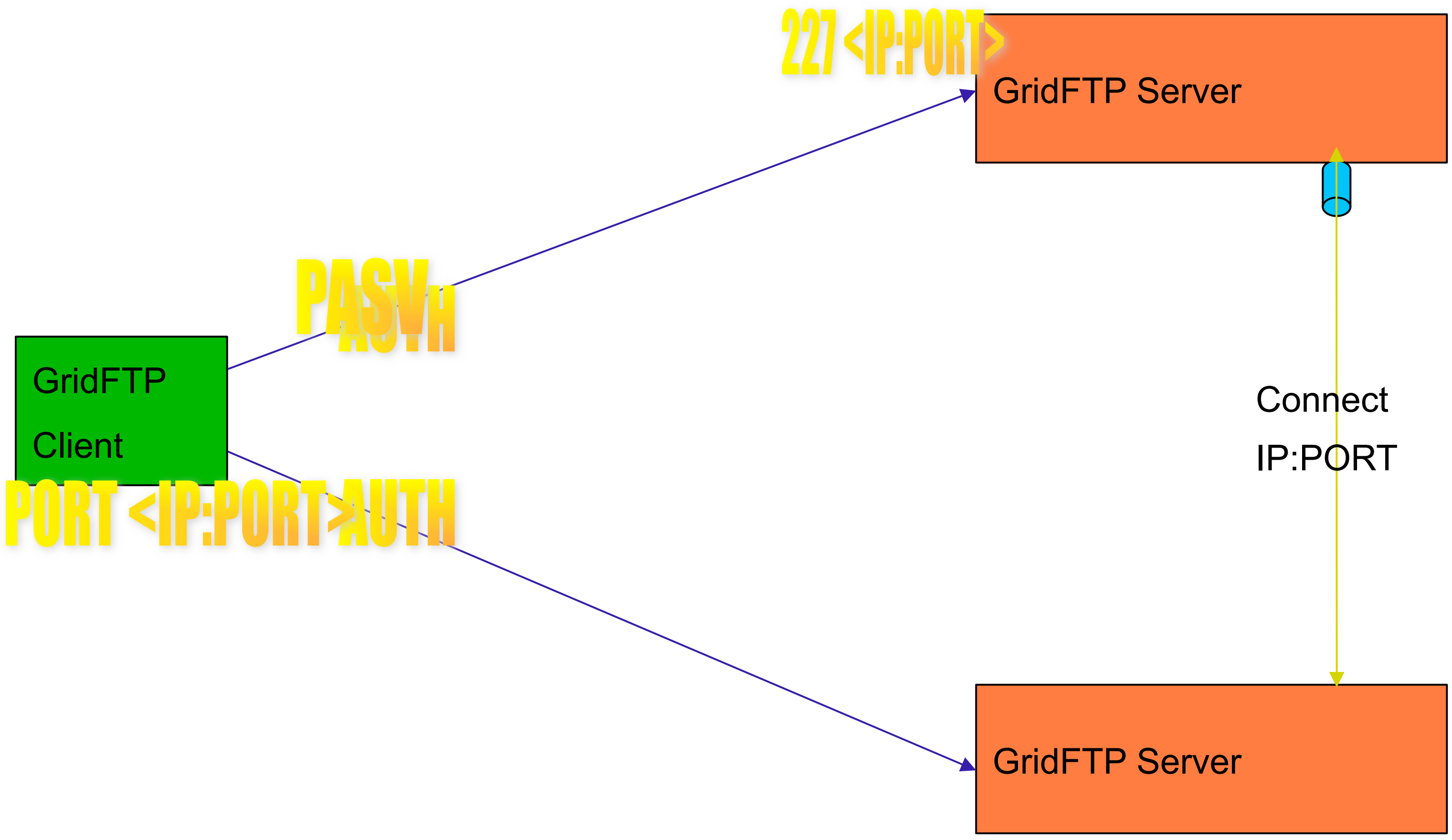
- Third party transfer
  - Client initiates data transfer between 2 servers
  - Client forms CC with 2 servers.
  - Information is routed through the client to establish DC between the two servers.
  - Data flows directly between servers
  - Client is notified by each server SPI when the transfer is complete



Control Channel Establishment

- Server listens on a well-known port (2811)
- Client form a TCP Connection to server
- 220 banner message
- Authentication
  - Anonymous
  - Clear text USER <username>/PASS <pw>
  - Base 64 encoded GSI handshake
- 230 Accepted/530 Rejected

Data Channel Establishment



227 <IP:PORT>

GridFTP Server

PASV

AUTH

GridFTP

Client

PORT <IP:PORT>

AUTH

Connect

IP:PORT

GridFTP Server

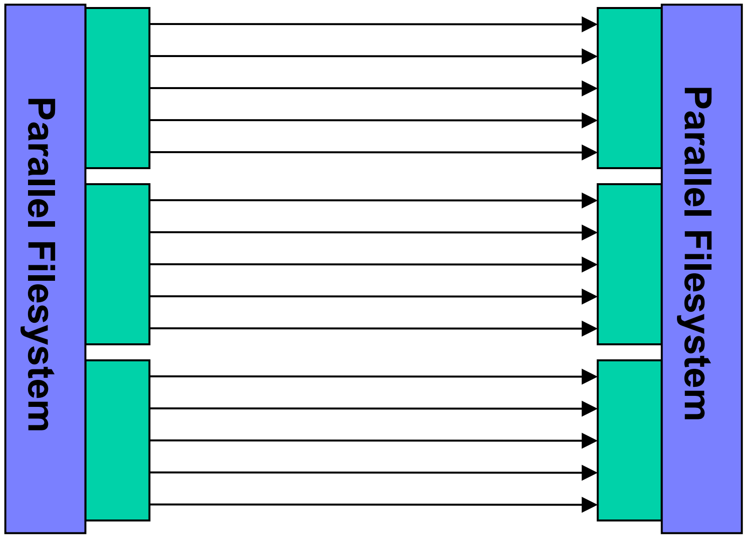
Data Channel Protocols

- MODE Command
  - Allows the client to select the data channel protocol
- MODE S
  - Stream mode, no framing
  - Legacy RFC959
- MODE E
  - GridFTP extension
  - Parallel TCP streams
  - Data channel caching

Descriptor (8 bits)	Size (64 bits)	Offset (64 bits)

Cluster-to-Cluster transfers

- Multiple nodes work together as a single logical GridFTP server
- Multiple nodes are used to transfer data into/ out of the cluster
  - Each node reads/writes only pieces they're responsible for
  - Head node coordinates transfers
- Multiple levels of parallelism
  - CPU, bus, NIC, disk etc.
  - Maximizes use of Gbit+ WANs

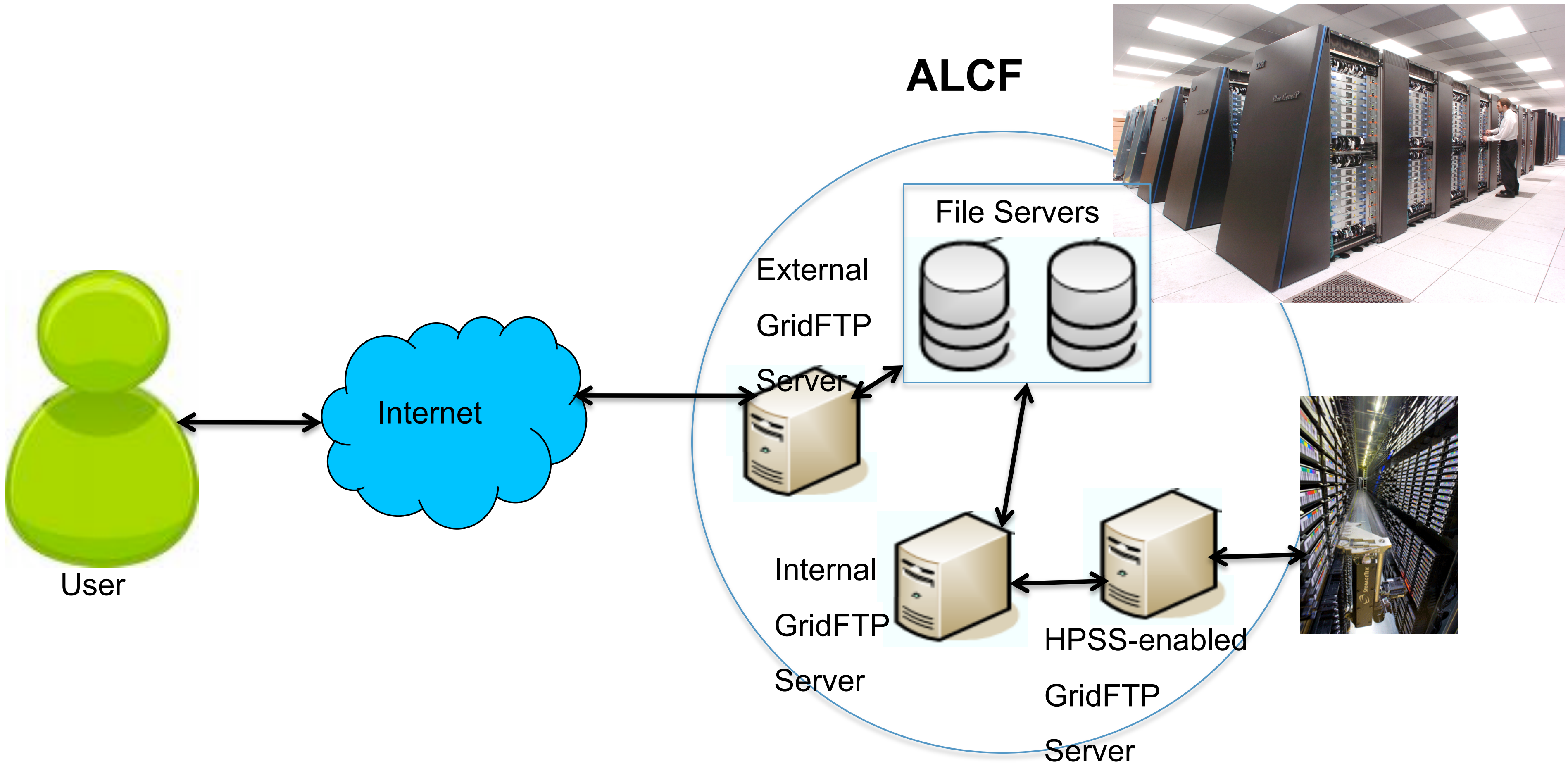


Striped Transfer
Fully utilizes bandwidth of Gb+ WAN using multiple nodes.

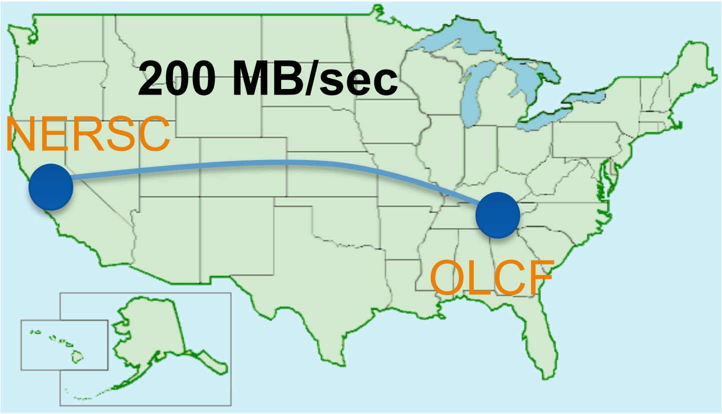
GridFTP in production

- Many Scientific communities rely on GridFTP
  - High Energy Physics - LHC computing Grid
  - Southern California Earthquake Center (SCEC), Earth Systems Grid (ESG), Relativistic Heavy Ion Collider (RHIC), European Space Agency, BBC use GridFTP for data movement
- GridFTP facilitates an average of more than 7 million data transfers every day

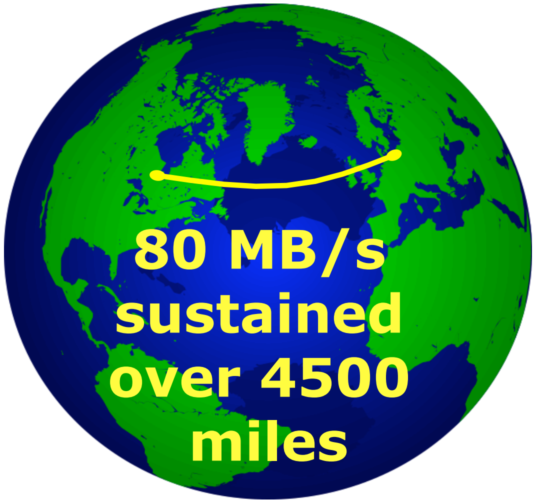
GridFTP in Production



GridFTP in production



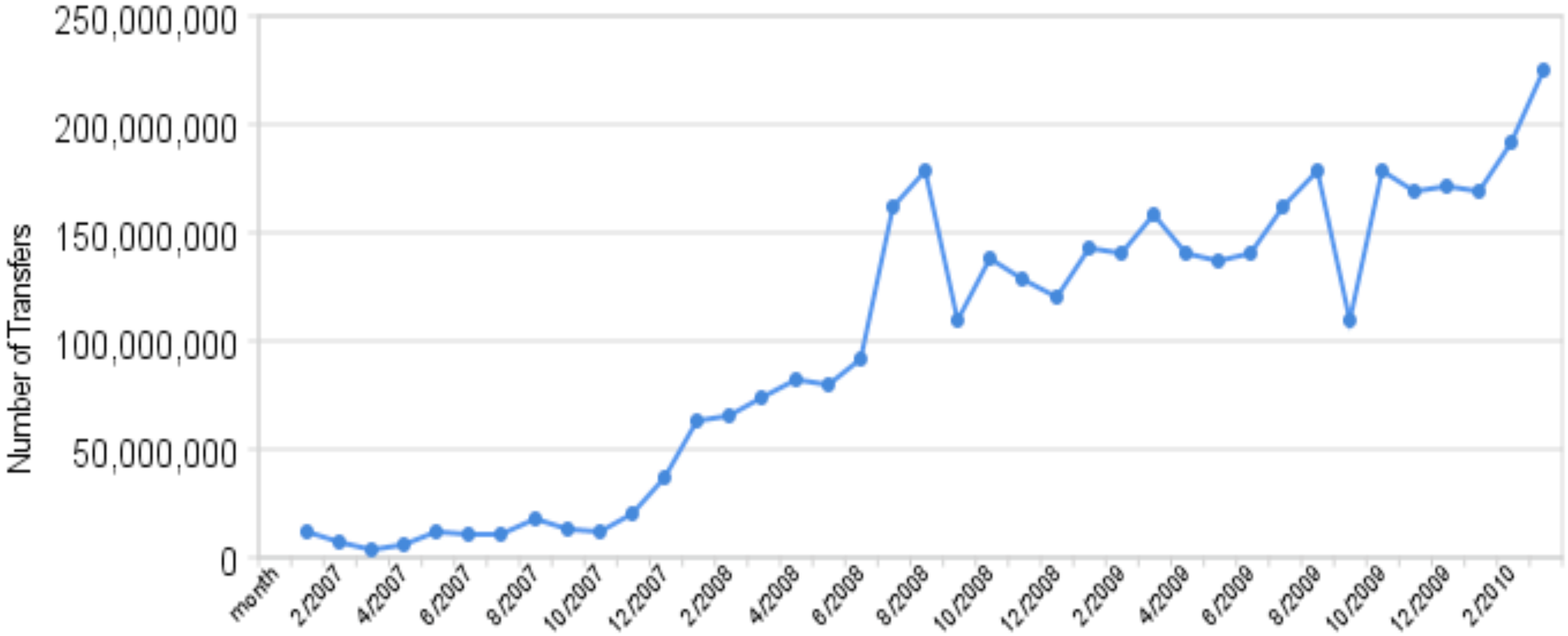
Move 40 terabyte (40 trillion bytes) from one DOE center (NERSC) to another (OLCF) in **under 3 days** rather than **several months**



1.5 terabyte moved from University of Wisconsin, Milwaukee to Hannover, Germany at a sustained rate of 80 megabyte/sec

GridFTP Usage

Monthly Totals* of GridFTP File Transfers



*for those "reporting"

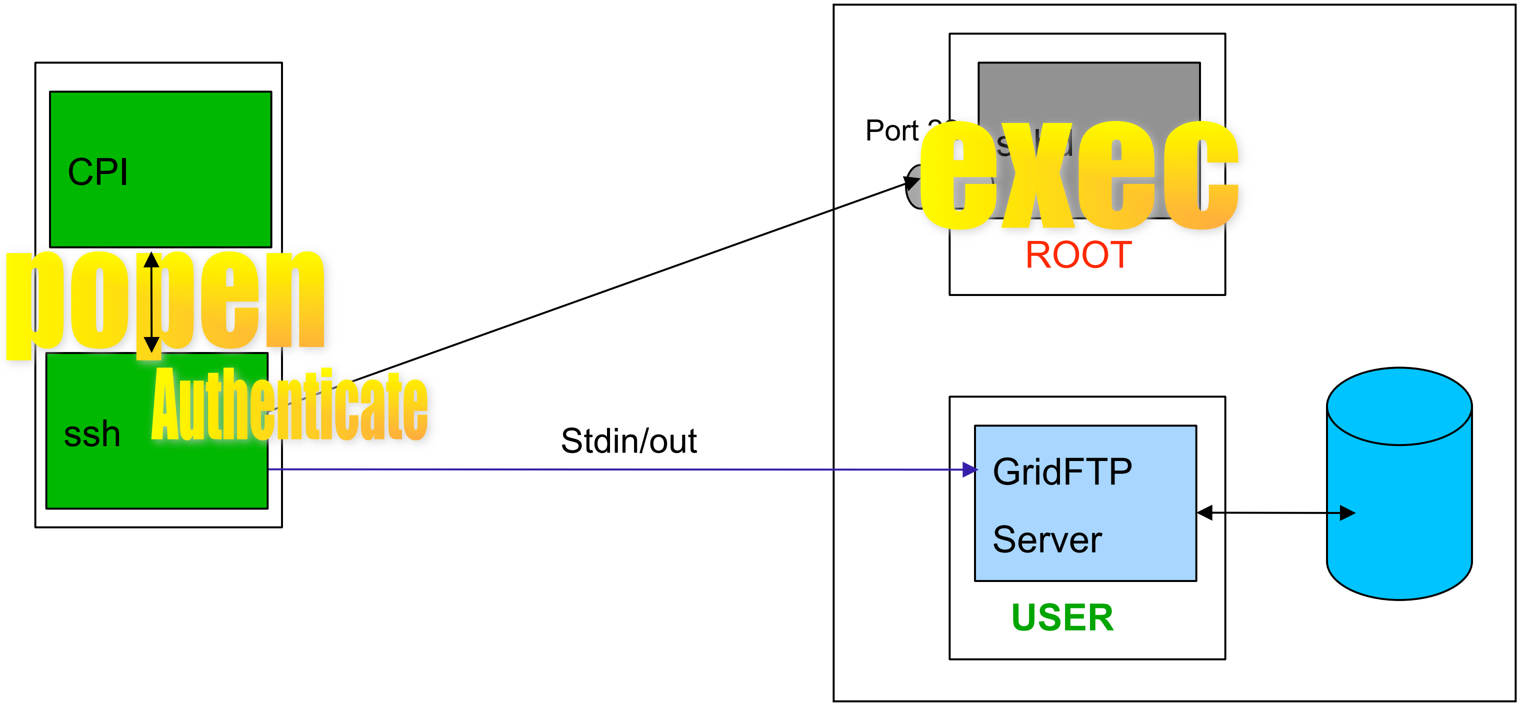
GSI

- Based on asymmetric cryptography
  - Private and Public Key - allows for two entities to authenticate with minimal cross-organizational support
- Certificates - Central concept in GSI
  - Information vital to identifying and authenticating user/service
  - Distinguished Name – unique Grid id for user/service
  - "/DC=org/DC=doegrids/OU=People/CN=Raj Kettimuthu 227852"
- Certificate Authority (CA)
  - Trusted 3rd party that confirms identity
- Host credential
  - Long term credential
- User credential
  - Passphrase protected

Security

- GridFTP provides strong security using GSI
- Protection vs. Ease of use
  - GSI and CAs were hard for many users
- Speed vs. protection
  - Users area happy with a minimal amount of data channel protection
- GridFTP over SSH
  - A big win for many users

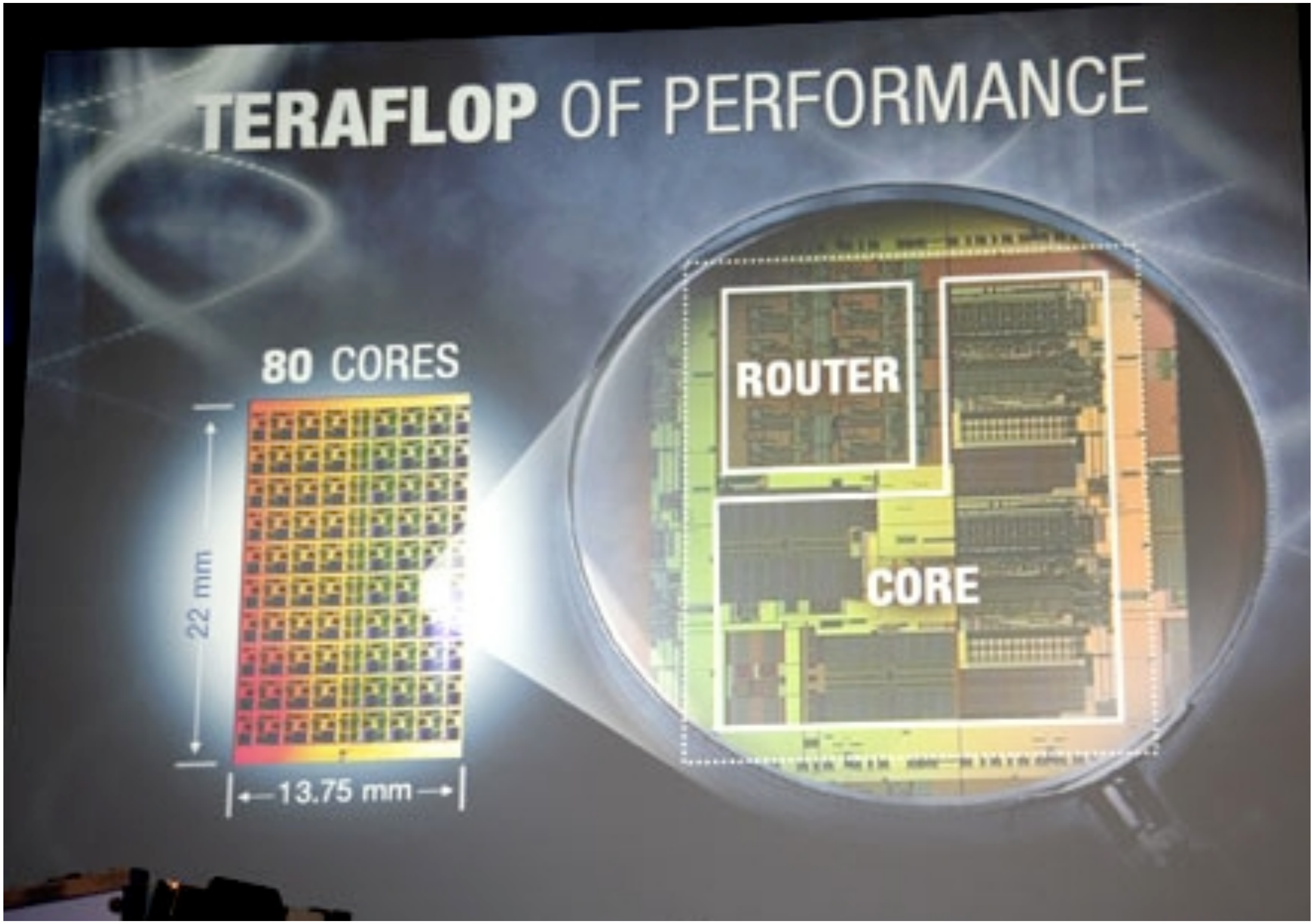
sshftp:// Interactions



Challenges

- Past success
  - Standard – big selling point for adoption
  - Throughput – GridFTP was sold on speed
  - Robustness – has to work all the time
  - Secure – data channel security
- Current and future
  - Extensible
  - Reliable
  - Scalable

Harnessing Multicore Architecture



Affinity

- Interrupt affinity
  - Interrupt processing done by processor to which the interrupt is physically bound
- Thread affinity
  - Application thread bound to processor where Interrupt processing of network traffic occurs.
- Memory affinity
  - Memory used by an application thread is allocated on the memory bank with the lowest access latency

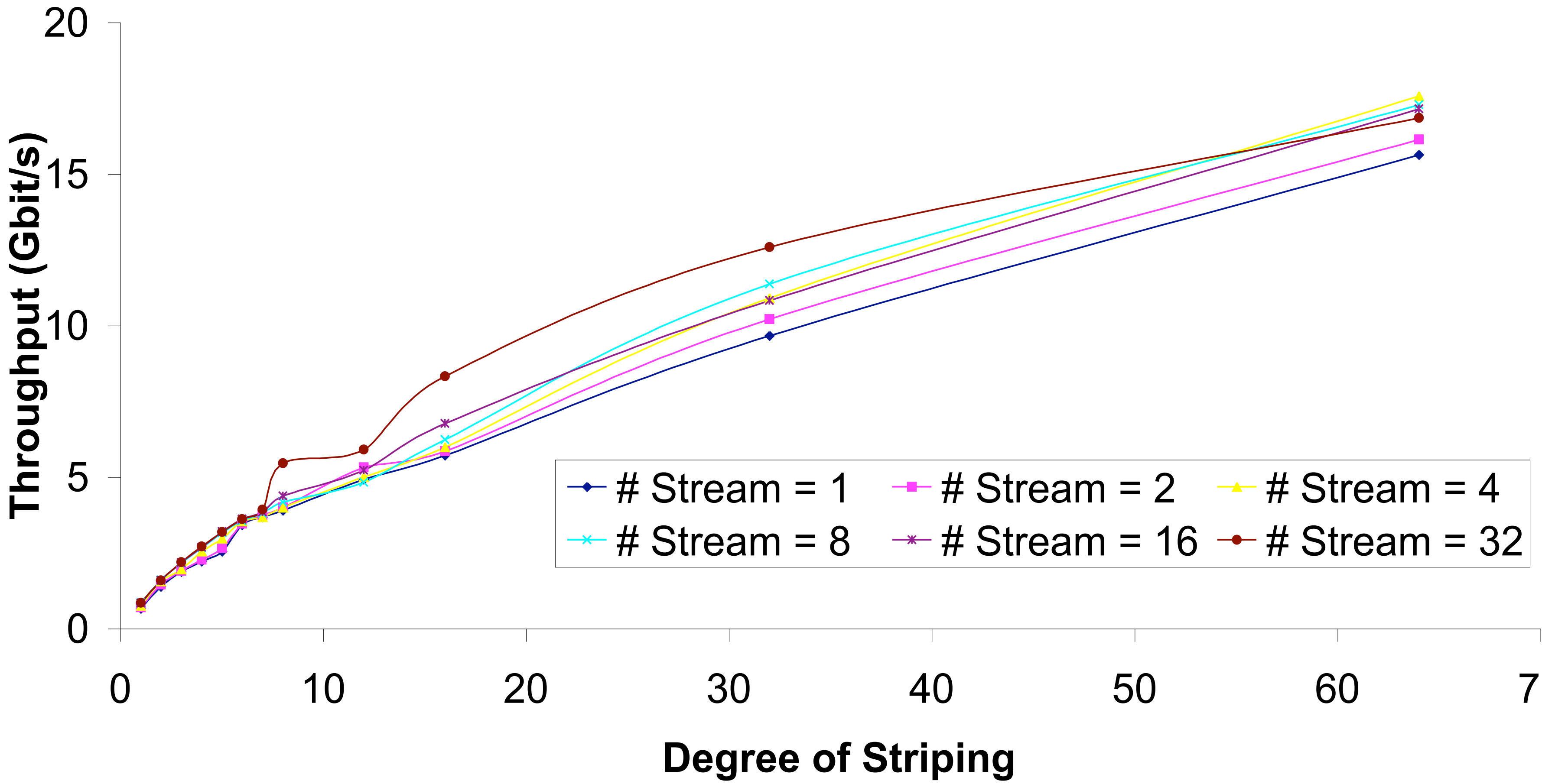
Dedicated Transfer Nodes

- Provides lot of compute power to drive transfers
- Most modern supercomputer architecture - parallel file systems optimized for high-performance local access
  - Typically massively parallel local access
  - Large collections of individual compute nodes accessing at the same time
  - Difficult to obtain optimal file system performance with few dedicated nodes

the globus alliance
www.globus.org

Example

- Disk transfer between Urbana, IL and San Diego, CA



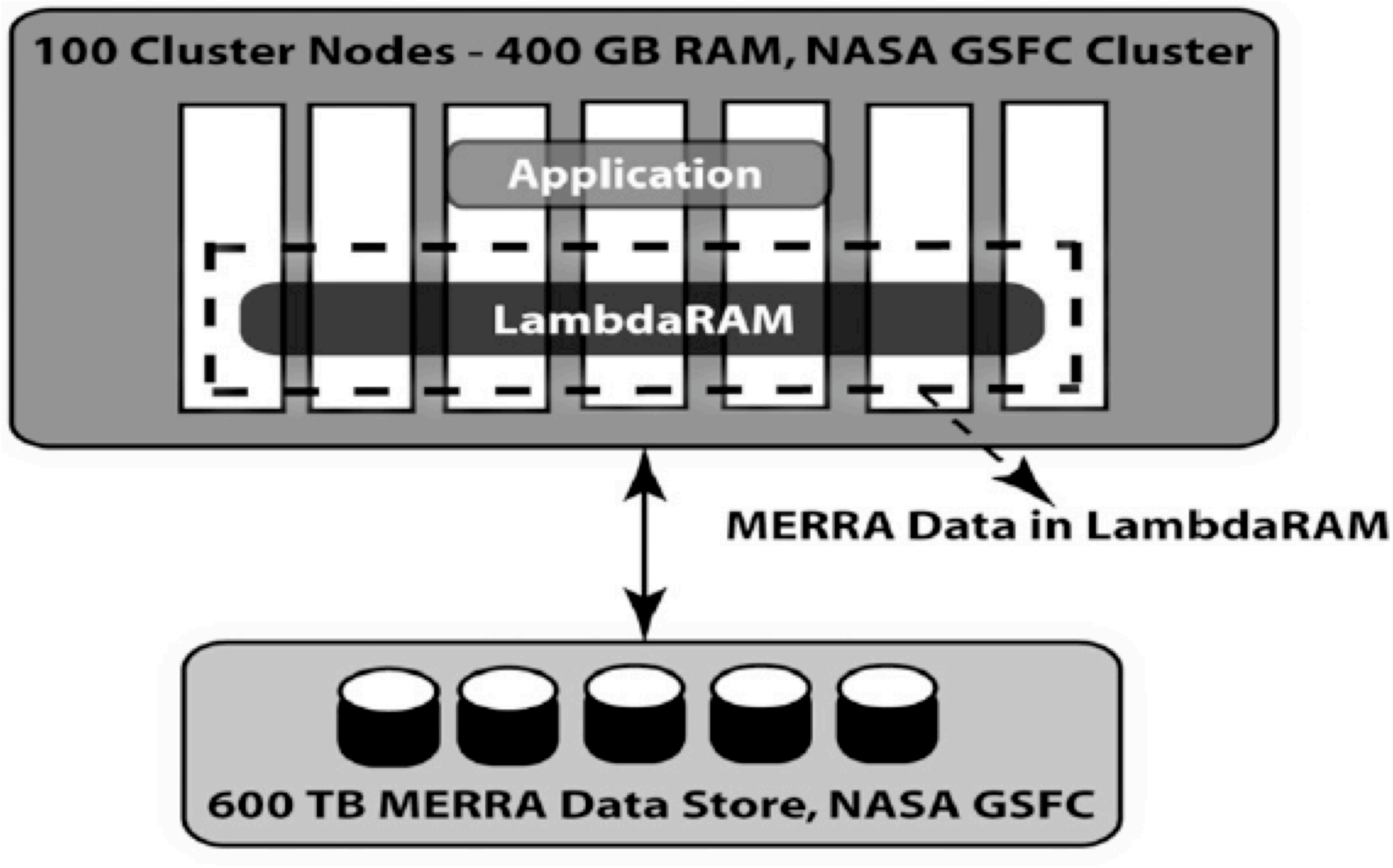
Functional Partitioning

- We have to "functionally partition" the compute node
  - Have cores that aid with network transfers - essentially bringing the "network into the job allocation."
  - More transfer nodes – better file system performance
  - Also help with direct streaming of data to the wide area - no need to the hit the scratch parallel file system

Efficient file systems

- To get 40 Gbit/s or 100 Gbit/s end-to-end, new storage techniques are needed
- Storage methods likely not POSIX complaint
- Interface with GridFTP
- Smart buffering in GridFTP may be warranted
  - Read/write in huge chunks
  - Reduce number of disk accesses

Distributed Memory Store

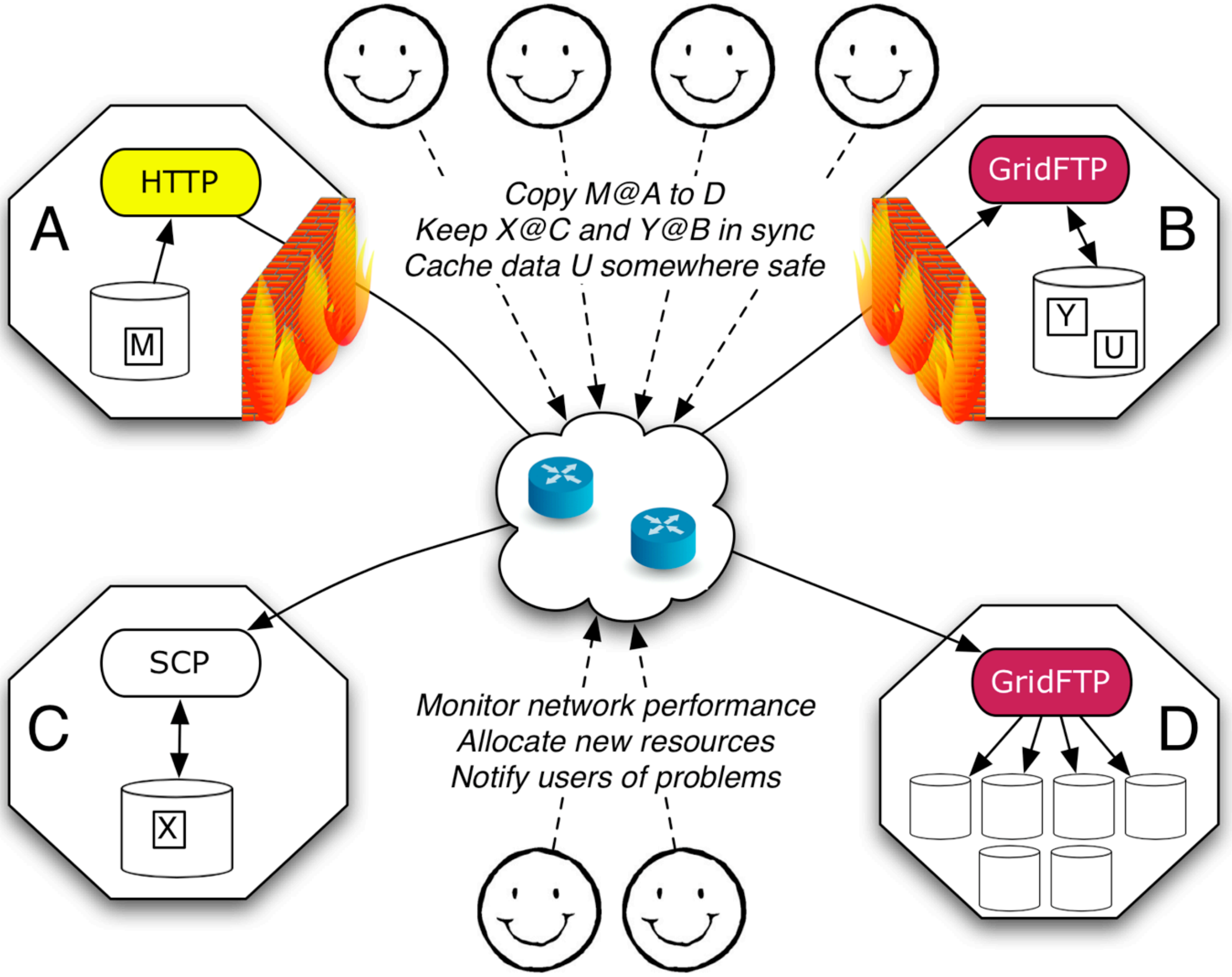

100 Cluster Nodes - 400 GB RAM, NASA GSFC Cluster
Application
LambdaRAM
MERRA Data in LambdaRAM
600 TB MERRA Data Store, NASA GSFC

400 GB LambdaRAM encompassing a Single Cluster

New Transport Protocols

- TCP has limitations in high-bandwidth and high-latency networks
  - With parallel streams, performance is acceptable in 10 Gbit/s links
- At 40 Gbit/s and beyond, new transport protocols necessary
  - RDMA based
  - Infiniband over WAN
- GridFTP needs to interface with these new protocols

Globus.org – hosted data movement service



Copy M@A to D
Keep X@C and Y@B in sync
Cache data U somewhere safe

Monitor network performance
Allocate new resources
Notify users of problems

Applying SaaS technique

- Service: Built as scale-out web application
  - Hosted on Amazon Web Services
  - Fire and forget
    - Less user interaction
    - Email notifications
  - Failure handling
    - Automatic retries
  - Familiar user interfaces
  - Technology interactions requiring no special expertise
  - No software to install

Globus.org

- Enable users to focus on domain-specific work
  - Manage technology failures
  - Notifications of interesting events
  - Provide users with enough information to resolve problems
- Ease the infrastructure providers' support burden
  - Hosted and supported by Globus team

More Information at http://www.gridftp.org http://www.globus.org/service/

Questions